SSTUT at NTCIR-4 Web task

Yinghui Xu Kyoji Umemura

Software System Lab. (Umemura Lab)
Information and Computer Science Dept.
Toyohashi University of Technology

June 3, 2004

Web Searching
Using term entropy on Virtual Document and Query Independent Importance

- Is the page itself adequate for Web IR ?
  - No. Page ‡ Document.
  - Page = textual page content + virtual document (VD).
- Does the term in query convey the same importance?
  - Usually not. Weighting query term may be helpful.
- What does linkage information of Web pages tell us?
  - Link analysis has been a good searching function for ranking web resources.

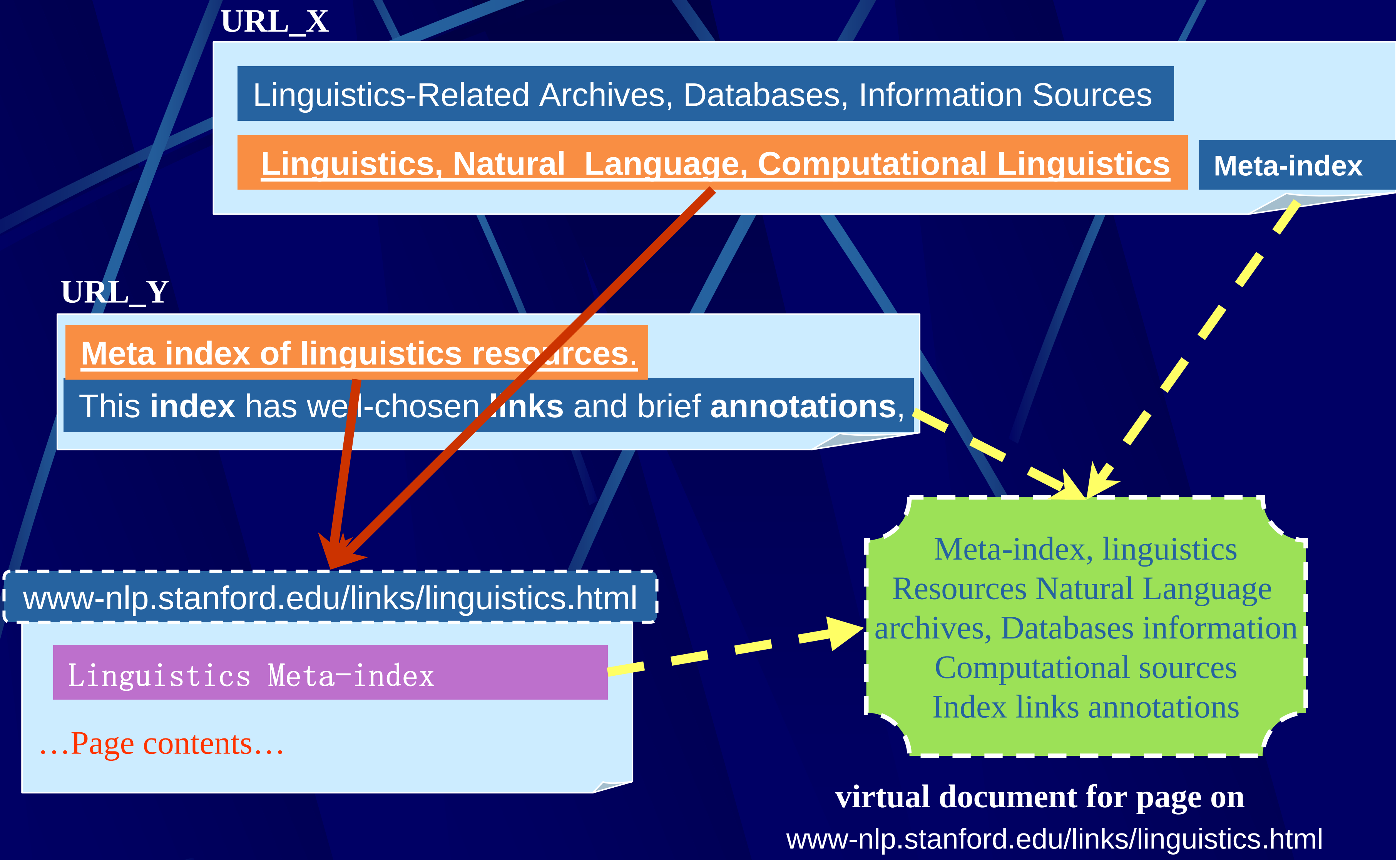
Our interests

- Feasible augmentation of general relevance ranking scheme through weighting query terms for Web IR.
- Effectiveness of information of VD on boosting the precision of general page content searching.
- Functionality of link analysis

Our Approach

- Weight query term based on term entropy in virtual document collection space and then introduced into general OKAPI model.
- Combining the relevance ranking score obtained through performing searching on both page content and page's virtual document.
- Proposing a literal matching aided link analysis model.

Sample Show of VD



URL_X

Linguistics-Related Archives, Databases, Information Sources

Linguistics, Natural Language, Computational Linguistics

Meta-index

URL_Y

Meta index of linguistics resources.

This **index** has well-chosen **links** and brief **annotations**,

www-nlp.stanford.edu/links/linguistics.html

Linguistics Meta-index

…Page contents…

Meta-index, linguistics
Resources Natural Language
archives, Databases information
Computational sources
Index links annotations

virtual document for page on
www-nlp.stanford.edu/links/linguistics.html

A diagram showing definition virtual document in our approach.

Definition of VD

- Comprised of the expanded anchor text from pages that point to him and some important words on the page itself.

$AnchorText(i, j)$: *set of terms appears in and around anchor of the link from i to j.*

$BodyText(j)$:

$$\begin{cases} \text{set of terms appearing in the "title" tag.} \\ \text{set of terms appearing in the meta tag.} \\ \text{set of terms appearing in the "H1, H2" tag.} \end{cases}$$

$VD(j)$: *set of terms in virtual document j.*

$$VD(j) = \left(\bigcup_i AnchorText(i, j) \right) \cup BodyText(j)$$

Assumption on VD

- Characteristic of VD:
  - Objective impression on page from others;
  - Subjective presentations of page author’s motivation.

- We assume:
  - VD is the representative information resources for Web pages.
  - VD is a good approximation of the type of summarization presented by users to search system in most queries.

Functionality of VD

- Allowing set up different weighting scheme and performing separate relevance ranking calculation.
- Predicting the query term importance.
- Providing the representative summarization of Web pages for deciding the transition probability in our proposed link analysis model.

Ranker – relevance ranking

- BASE - OKAPI’s BM25

$$SIM(Q,d)=\sum_{w\in Q}\frac{tf}{tf+0.5+1.5*{dl}/{ave\_dl}}\times\frac{\log_2(0.5+N/df)}{\log_2(1.0+\log_2(N))}$$

- QTIBRF
  - Query term importance based ranking function

$$SIM(Q,d)=\sum_{w\in Q}\boxed{VDTW(w)}\times\frac{tf}{tf+0.5+1.5*{dl}/{ave\_dl}}\times\frac{\log_2(0.5+N/df)}{\log_2(1.0+\log_2(N))}$$

- SMRF – score merging ranking function

$$FinalScore(p_i)=SIM(Q,VD(p_i))+\lambda SIM(Q,AD(p_i))$$

$$\lambda=0.114$$

Query term weighting in QTIRBF

- Query term are weighted by its entropy on virtual document collection space.

$$VDTF\left(w, j\right) = \#\left\{w \mid w \in VD\left(j\right)\right\}$$

$$P\left(w, j\right) = VDTF\left(w, j\right) \Big/ \sum_{k=1}^{N} VDTF\left(w, k\right)$$

$$VDET\left(w\right) = -\sum_{j=1}^{N} P\left(w, j\right) \log_N P\left(w, j\right)$$

⇩

$$VDTW\left(w\right) = 1 - VDET\left(w\right)$$

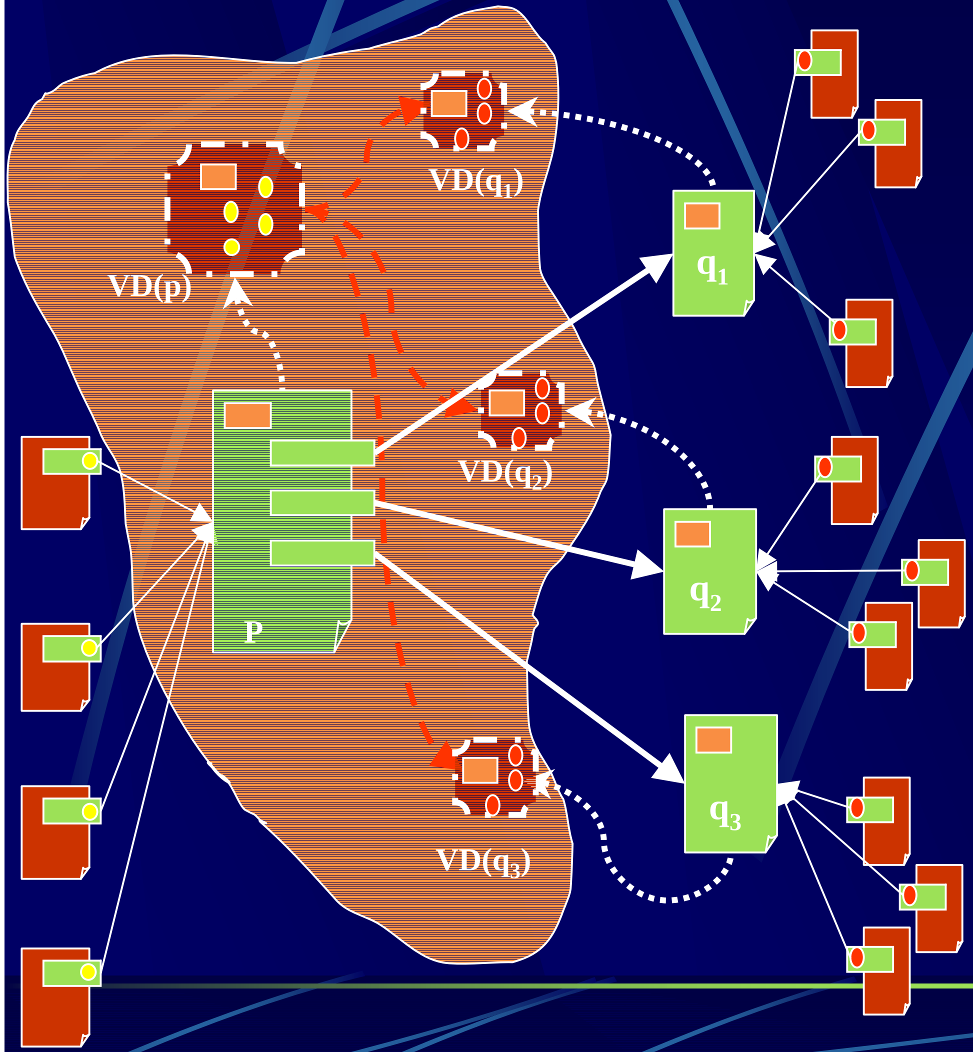
LinkAnalyzer
- Literal Matching aided link analysis

- What we hold:
  - Inbounds links from pages with similar theme to our own have larger influence on PageRank than links from unrelated pages
- Our approach:
  - Combine the evidence from both content and link structure into the link analysis method
  - Modify the underlying Markov process by giving different weights to different outgoing links from a page.

Assumption

- User would like to choose the relevant target that they picture in their mind.
- Searching is a process to approach a desired outcome of user gradually. Accordingly, user’s mind are somewhat consistent in searching path.

Diagram of LMALA



- *TranOdds(P→q$_k$)*
  - *prob(VD(q$_k$)|P)*
    - Measure how likely the VD of the activate target page can be generated by the page being viewed
  - $\sum_{w \in (VD(q_k) \cap VD(p))} prob\left(\frac{w}{p}\right)$
    - indicate the dependent degree of the two connected VD. Measure user 's mind consistency

Computation Model

- Based on calculated values that indicate transition likelihood for all possible connections on a page, we assign the transition probability to them and regard them as the link weight in the Markov chain.

$$\lambda = 0.85$$

$$\gamma = 0.7$$

$$PR(j) = (1-\lambda)1/N + \lambda \sum_{i \in B_j} PR(i)\, prob(i \to j)$$

$$prob(i \to j) = \begin{cases} \gamma \times \dfrac{TranOdds(i \to j)}{\sum_{k \in F(i)} TranOdds(i \to k)}, & Liter(link(i,k)) = 1 \\ (1-\gamma) \times \left(1 \big/ \left(\# F(i) - LiterLink(i)\right)\right), & otherwise \end{cases}$$

The condition represent whether the link between *i* and *k* has relevant literal information or not.

Rank adjuster

- Model 1. (RA1)

$$FScore(P_i) = SMRF(P_i) + \lambda \times \frac{\log(LMALA(P_i) * N)}{\log(1.8)} \qquad \lambda = 0.1$$

- Model 2. (RA2)

$$FScore = SMRF(P_i) - \lambda \times \frac{\tau_1(P_i) + \tau_2(P_i)}{|\tau_1(P_i) - \tau_2(P_i) + 1|} \qquad \lambda = 0.08$$

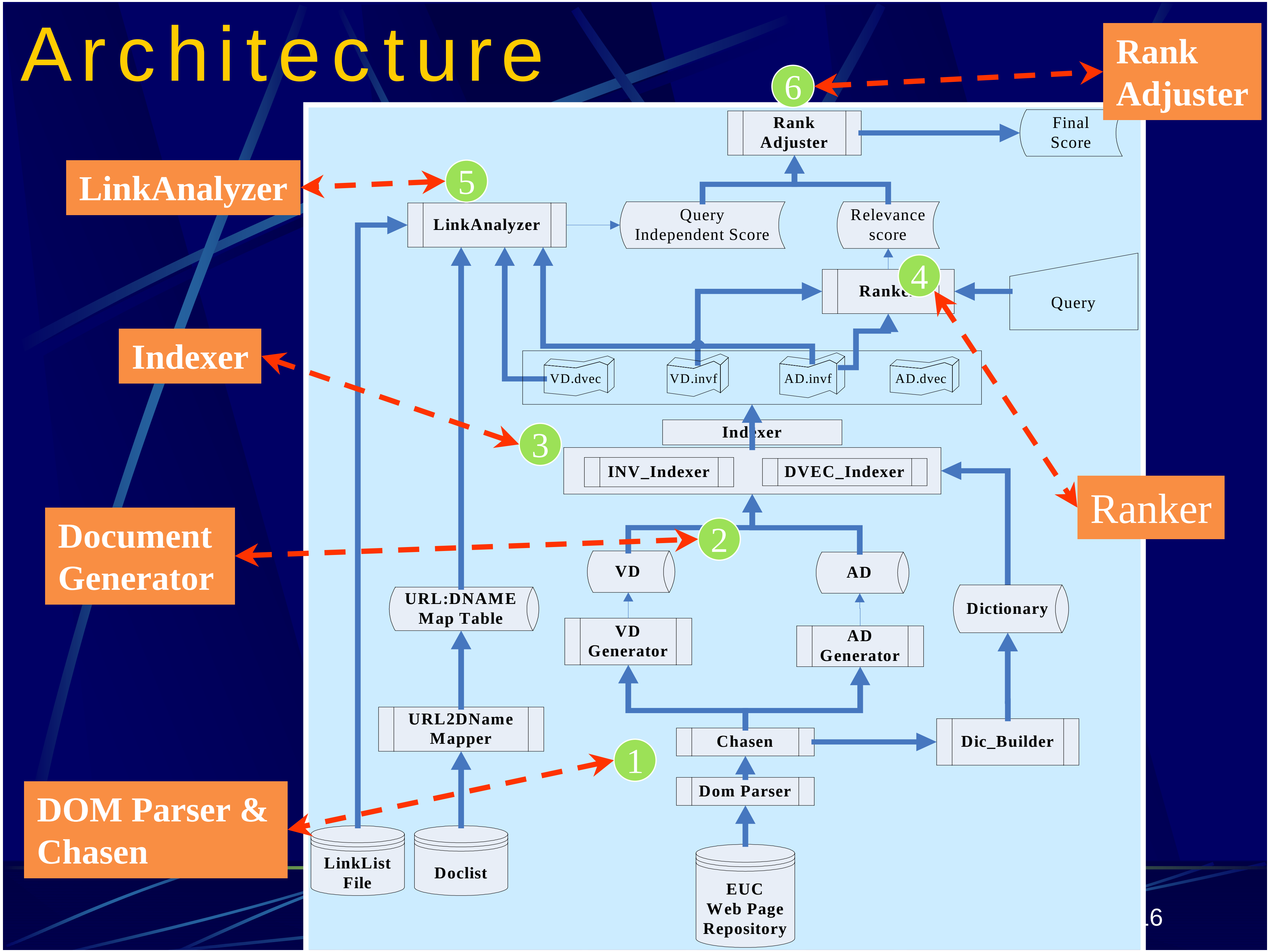
$R:$ *return document sets for a given query*

$\tau_1:$ *document in R sort by SMRF score*

$\tau_2:$ *document in R sort by LMALA score*

$\tau_k(i):$ *rank of i in* τ_k

Architecture

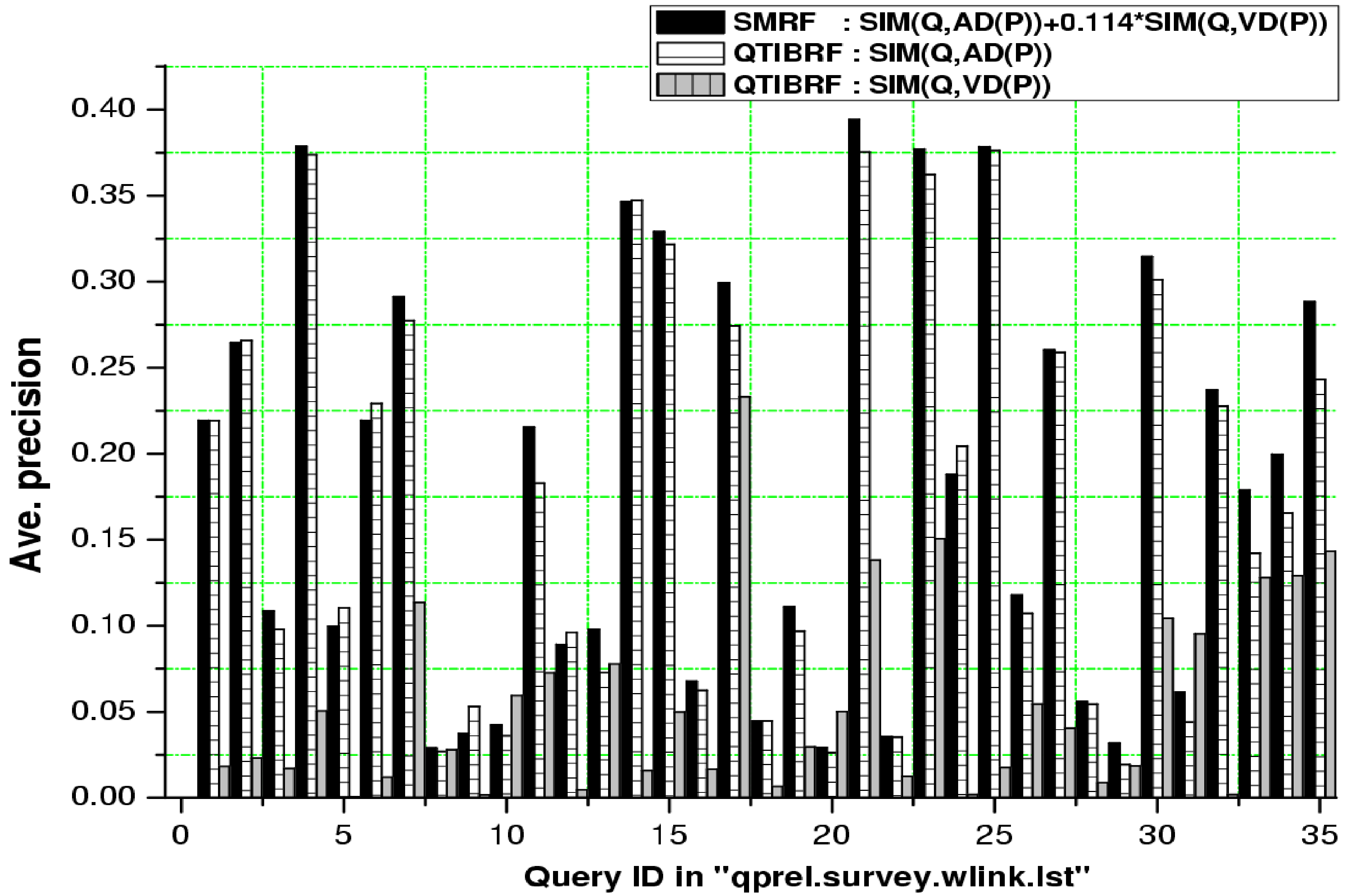


Experiment results
- BASE vs. QTIBRF

		Virtual document (VD)			Actual document (AD)		
	Topic	Ave. P	P@10	P@20	Ave.P	P@10	P@20
BASE	tt	0.0621	0.2738	0.2206	0.2052	0.4550	0.3931
QTIBRF	tt	0.0705	0.2850	0.2431	0.2127	0.4487	0.3850
BASE	desc	0.0579	0.2550	0.2038	0.1839	0.4300	0.3713
QTIBRF	desc	0.0641	0.2825	0.2306	0.1987	0.4225	0.3625

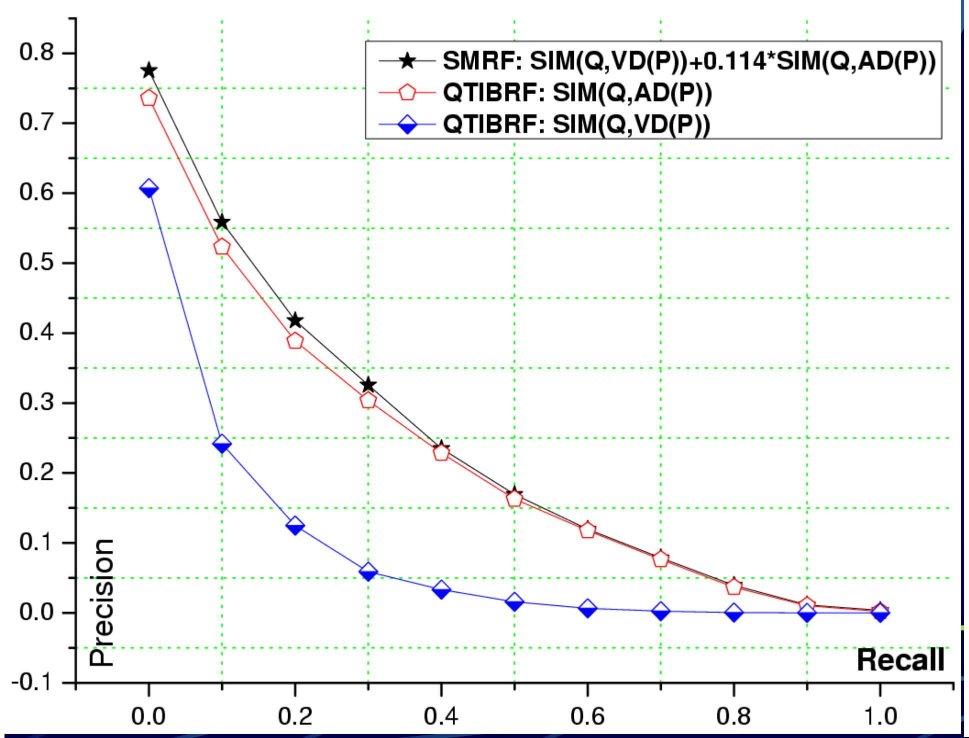
- QTIRBF got improvements of Ave. P on both VD and AD searching.
- QTIRBF is more adaptable for improving VD based searching

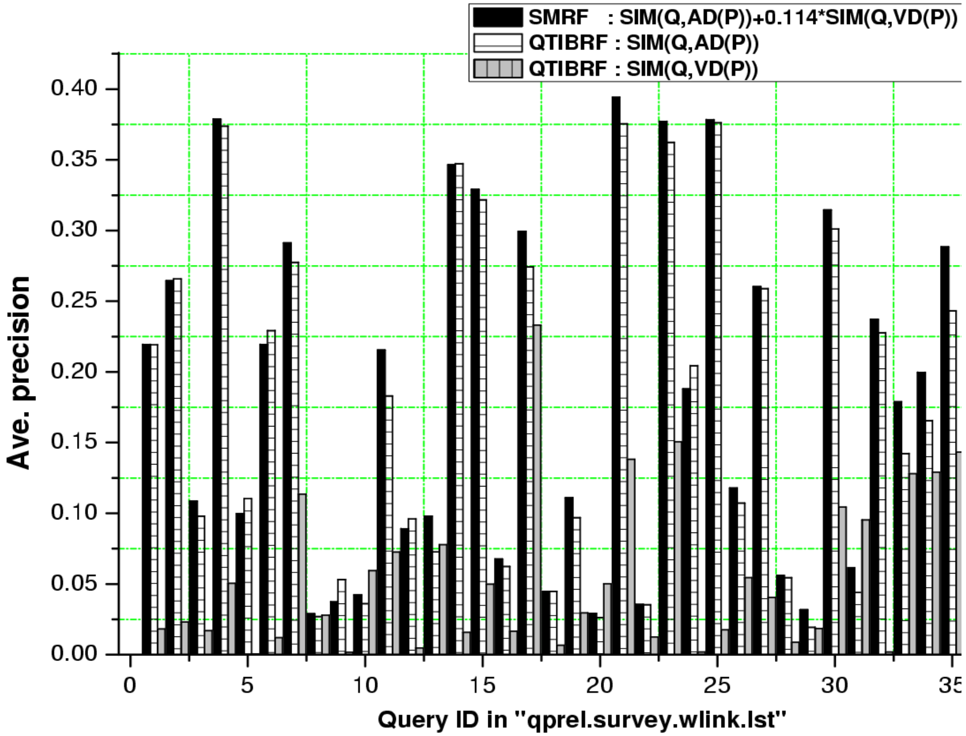
SMRF vs. QTIBRF



SMRF vs. QTIBRF

	Rank Fun.	Ave.P	P@10	P@20
VD only	QTIBRF	0.0705	0.2850	0.2431
AD only	QTIBRF	0.2127	0.4437	0.3750
VD+AD	SMRF	0.2208	0.4767	0.4184

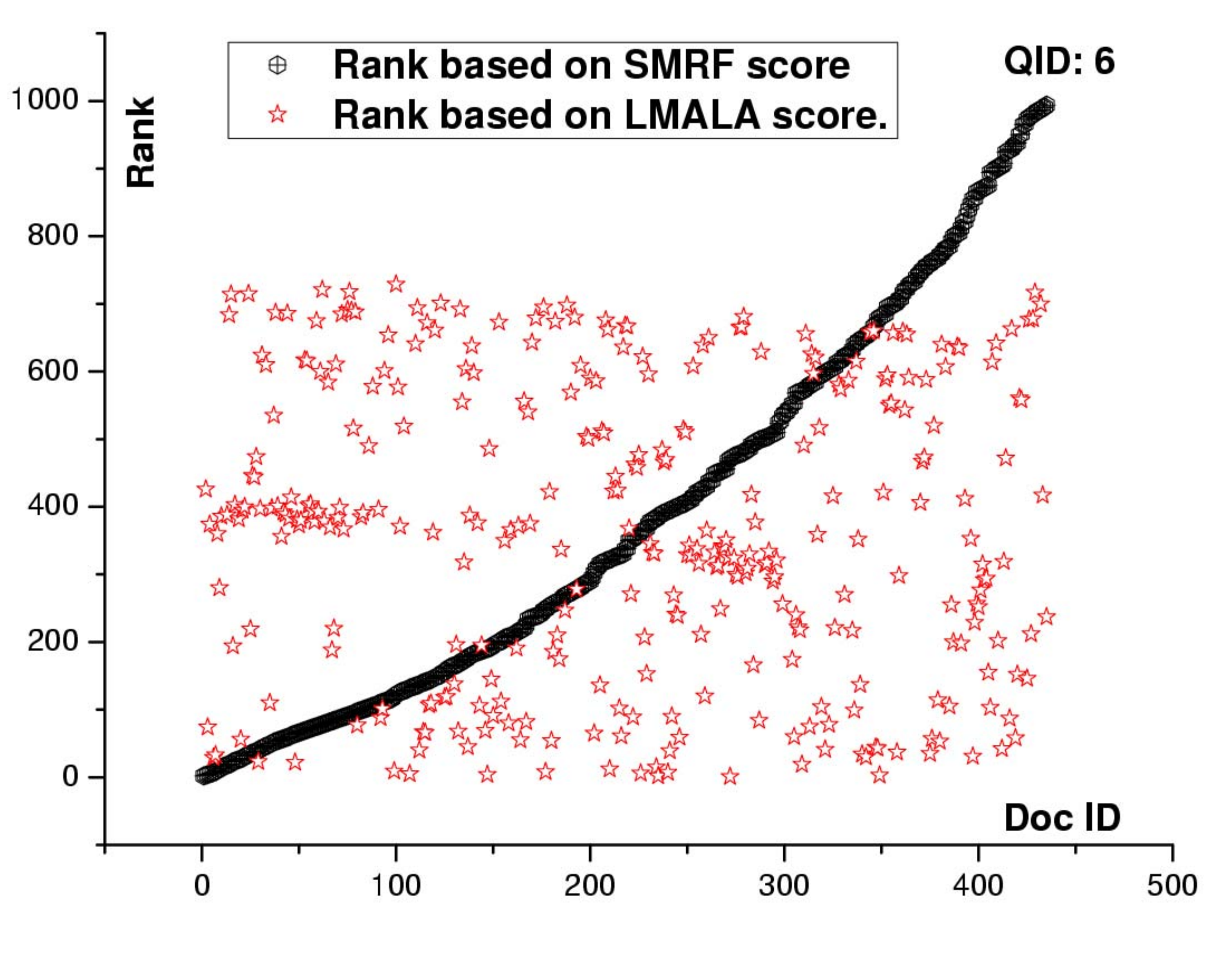




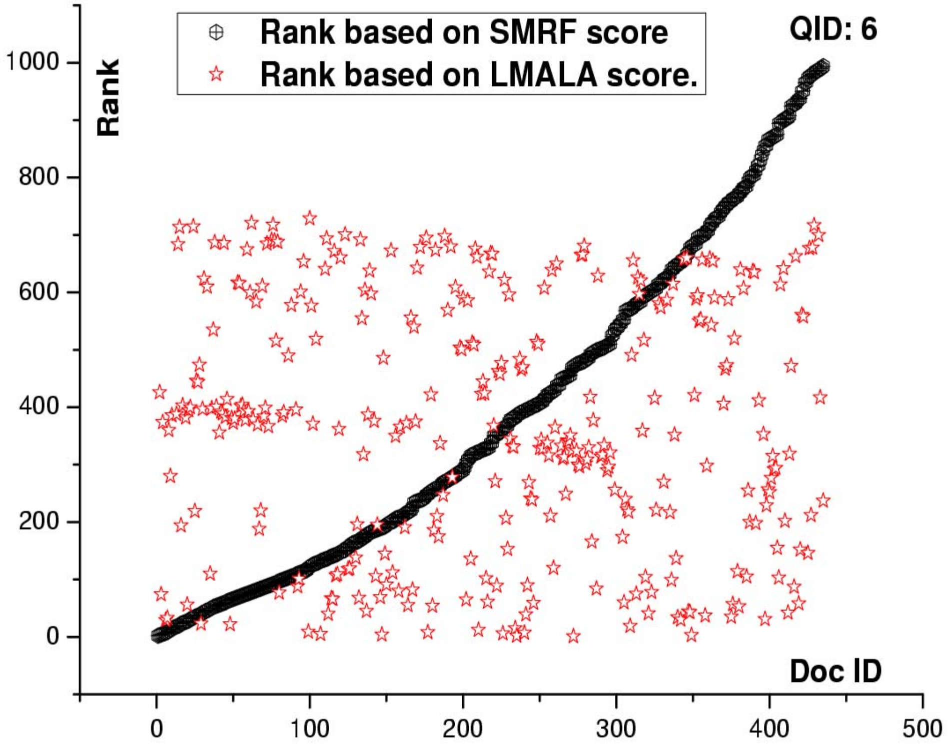
SMRF vs. RA1 and RA2

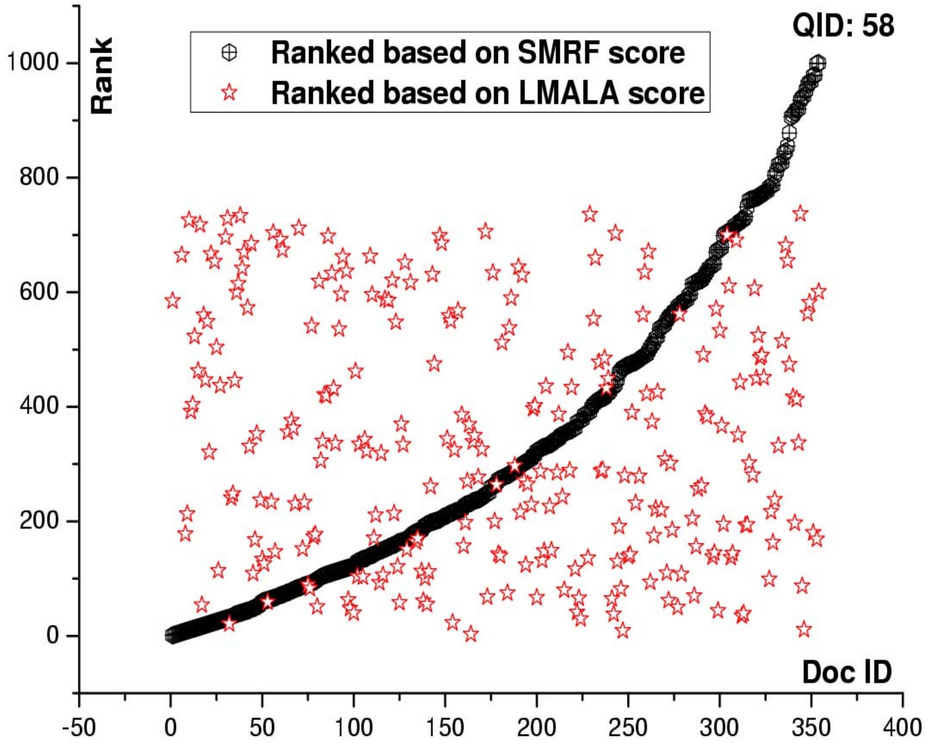
		SMRF	RA1	RA2
Ave. P		0.1203	**0.1212**	0.1204
Recall	**0.0**	0.7036	0.7116	**0.7226**
	0.1	0.4157	**0.4246**	0.4143
	0.2	0.2576	**0.2577**	0.2557
	0.3	0.1751	**0.1759**	0.1740
Prec.	**@5**	0.4629	0.4457	0.4629
	@10	0.4000	0.3943	**0.4057**
	@20	0.3529	0.3514	**0.3543**
	@30	0.3314	0.3286	03343

Rank comparison of relevant file



Rank comparison of relevant file





Conclusion

- Weighting query term through entropy on VD space improves searching results.
- It indicates that the system which makes used of Web structure, such as anchor, title, will perform better than the content-only system without considering them.
- No clear improvements obtained by combining query independent score using our proposed link analysis model, but indicate the potential ability on improving searching results.

Thank you